Statistics 2000, Section 001, Midterm 1 (185 Points)

Friday, February 11, 2011

Part I: Text Answers

Your Name: _____

-2 for each calculation error (or no final result)

Question 1: z–Scores and Normal Distributions (**50 Points**)

The temperature at any random location in a kiln used for manufacturing bricks is normally distributed with a mean of 1000°F and a standard deviation of 50°F.

Show your work!

1. (**10 Points**) What is the z–score that relates to 1075°F?

$z = \frac{1075 - 1000}{50} = 1.50$ (10)

2. (**10 Points**) A z–score of −2.75 relates to a temperature of 862.50 °F?

$x = -2.75 \cdot 50 + 1000 = 862.50$ (10)

3. (**10 Points**) If bricks are fired at a temperature above 1125°F, they will crack and must be discarded. If the bricks are placed randomly throughout the kiln, what is the percentage of bricks that crack during the firing process?

$z = \frac{1125 - 1000}{50} = 2.50$ (5)

area to the right of 2.50 = area to the left of −2.50 = 0.0062 = 0.62% (5)

4. (**10 Points**) When glazed bricks are put in the oven, if the temperature is below 900°F, they will discolor. If the bricks are placed randomly throughout the kiln, what percentage of glazed bricks will discolor?

$z = \frac{900 - 1000}{50} = -2.00$ (5)

area to the left of −2.00 = 0.0228 = 2.28% (5)

5. (**10 Points**) Suppose the kiln can hold 10,000 bricks at a time. When completely filled, bricks at the 1,000 hottest locations will be exposed to temperatures of 1064 °F (and higher).

to be in the top 10%, we need an area of 0.90 (90%) to the left. (3)

This is the case for $z \approx 1.28$ (3) (working with 1.29 is fine);

$\rightsquigarrow$ temperature = $1.28 \cdot 50 + 1000 = 1064$ (4)

based on: HW2, Exercise 1, Q1.5 & 6

Question 2: Boxplots (**45 Points**)

Longleaf pine trees: The Wade Tract in Thomas County, Georgia, is an old–growth forest of longleaf pine trees (*Pinus palustris*) that has survived in a relatively undisturbed state since before the settlement of the area by Europeans. A study collected data about 584 of these trees. One of the variables measured was the diameter at breast height (DBH). This is the diameter of the tree at 4.5 feet and the units are centimeters (cm). Only trees greater than 1.5 cm were sampled. Here are the diameters of a random sample of $n = 11$ of these trees. The data have been sorted from smallest to largest:

i	$x_{(i)}$	
1	2.3	← Min
2	2.7	
3	11.4	← Q_1
4	13.3	
5	18.3	
6	26.0	← M
7	32.6	
8	43.6	
9	47.2	← Q_3
10	52.2	
11	69.3	← Max

-2 for each calculation error (or no final result)

} [or $Q_1 = \frac{11.4+13.3}{2} = 12.35$]

} [or $Q_3 = \frac{43.6+47.2}{2} = 45.4$]

Show your work!

1. **(20 Points)** Calculate the values for the five number summary for DBH and clearly name each of these values (e.g., if variance is one of these numbers than indicate "variance = ..."). (1) each correct name (3) each correct value

   1.) Minimum Min = 2.3
   2.) First Quartile Q_1 = 11.4
   3.) Median M = 26.0
   4.) Third Quartile Q_3 = 47.2
   5.) Maximum Max = 69.3

2. **(5 Points)** First indicate the formula, and then calculate the IQR.

   $IQR = Q_3 - Q_1$ (2) $= 47.2 - 11.4$ (3) $= \underline{35.8}$

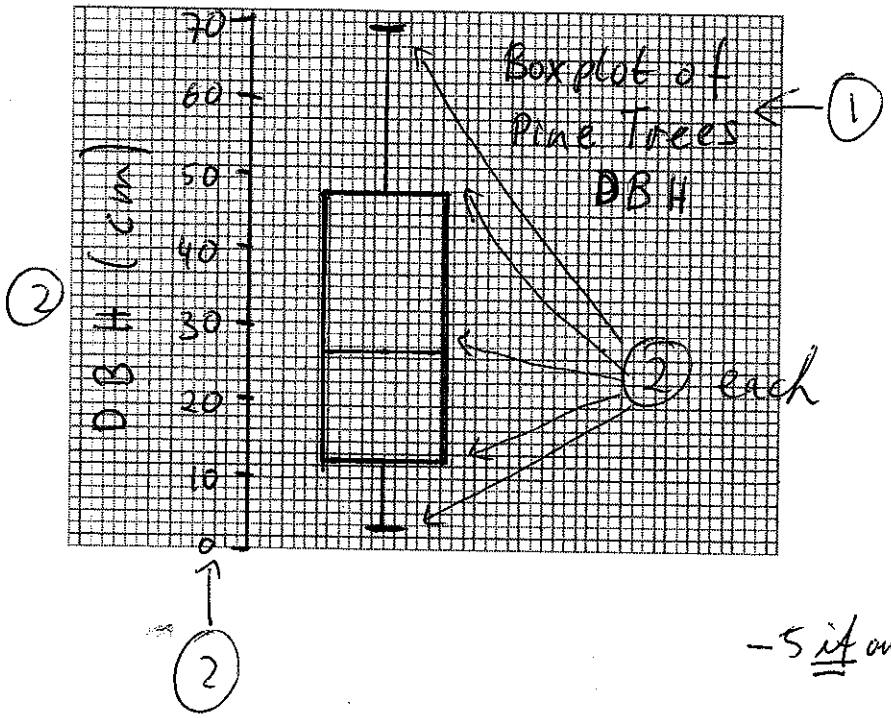
   [or $45.4 - 12.35 = \underline{33.05}$]

Note: An observation is a suspected outlier if it falls more than 1.5·IQR above the third quartile or below the first quartile. Here, 1.5·IQR = 1.5·35.8 = 53.7

3. **(15 Points)** Draw a boxplot for the 11 observations above from the DBH data set. Make sure to label your graph! Be careful with outliers (in case there are any).

$\Rightarrow Q_3 + 1.5 \cdot IQR = 47.2 + 53.7 = 100.9$
and $Q_1 - 1.5 \cdot IQR = 11.4 - 53.7 = -42.3$
$\Rightarrow$ there are no outliers in our data set!

(+4) bonus points for this calculation



(1) (2) (2) each

−5 if outlier(s) are drawn

4. **(5 Points)** Based on your boxplot, is the distribution of DBH (i) roughly symmetric, (ii) skewed to the right, or (iii) skewed to the left? **Just circle your answer.**

(5)

[Note that the distance from Q_3 to the Max is much bigger than the distance from Q_1 to the Min.]

Question 3: Regression (**30 Points**)

Data were obtained from the A&W Web site for the Total Fat in grams and the Protein content in grams for various items on their menu. Some summary statistics are also provided:

Item	Total Fat (grams)	Protein (grams)
Kid's Cheeseburger	24	23
Kid's Hamburger	22	21
Original Bacon Cheeseburger	33	27
Original Bacon Double Cheeseburger	48	45
Original Double Cheeseburger	42	40
Papa Burger	42	41

	y = Total Fat		x = Protein
Mean	35.167		32.833
Standard Deviation	10.591		10.362
Correlation		$r = 0.983$	

-2 for each calculation error (or no final result)

-2 if x, y flipped

The scatterplot (not reproduced here) shows that there is indeed a linear relationship between the two variables.

-1 if < 3 decimal digits (each time)

Work with 3 decimal digits (as above) and show your work!

1. (**20 Points**) Find the regression equation for predicting the Total Fat from the Protein content.

slope: $b_1 = r \cdot \frac{s_y}{s_x} = 0.983 \cdot \frac{10.591}{10.362} = 1.005$ (8)

y-intercept: $b_0 = \bar{y} - b_1 \bar{x} = 35.167 - 1.005 \cdot 32.833 = 2.170$ (8)

regression equation: $\hat{y} = 2.170 + 1.005x$ (4)

2. (**10 Points**) Using your regression equation, estimate the Total Fat for a menu item with a Protein content of 25 grams.

$\hat{y}$ (for 25) $= 2.170 + 1.005 \cdot 25 = 27.295$ (10)

Note: Statistics is not about punching numbers in a calculator, but about interpreting data. A result that makes no sense, but is left without interpretation, results in a "-4" point deduction. When you get an implausible result (such as -20 or 150), you must comment that something went wrong!

Statistics 2000, Section 001, Midterm 1 (185 Points)

Friday, February 11, 2011

Part II: Multiple Choice Questions

Your Name: _____

Question 4: Multiple Choice Questions **(60 Points)**

Mark your answer for each multiple choice question in the table below. There is only one correct answer for each question. Each correct answer is worth 4 points.

Question	(a)	(b)	(c)	(d)	Question	(a)	(b)	(c)	(d)
1	●	○	○	○	11	○	●	○	○
2	○	○	●	○	12	○	●	○	○
3	●	○	○	○	13	○	○	●	○
4	●	○	○	○	14	○	○	○	●
5	○	●	○	○	15	○	●	○	○
6	●	○	○	○					
7	●	○	○	○					
8	○	○	○	●					
9	●	○	○	○					
10	●	○	○	○					

based on:

- 5 questions based on "Online Quizzes" for Chapters 1 & 2
- 5 questions based on previous versions of Stat 2000, Midterm 1 (Fall 2009 - Fall 2010)

Stat 2000, Midterm 1, Question 4 — Solutions

1. (a) The closer the points in a scatterplot are to lying along a straight line, the closer the correlation is to $+1$ or -1. If the straight line has a positive slope (slopes up from left to right), the correlation is positive, and if the straight line has a negative slope (slopes down from left to right), the correlation is negative. In this case, the points in the plot fall roughly along a line that slopes up from left to right, so we expect the correlation to be positive and have a value closer to $+1$ than to 0.

2. (c) The correlation r measures association between two quantitative variables, and both variables are quantitative in this case. We would expect the correlation to be positive. The correlation can't be bigger than $+1$ and the correlation is unitless.

3. (a) Females always mate with males that are 0.50 years younger, so the relationship is perfectly linear, and as females get older so will the males they mate with, so the correlation is 1.

4. (a) Only (a) is correct since $b_1 = r\frac{s_y}{s_x}$, where b_1 is the slope, r is the correlation, and the ratio of the standard deviations $\frac{s_y}{s_x}$ is always positive. So, b_1 must have the same sign as r.

5. (b) The stem–and–leaf plot gives the individual observations, so it is possible to compute the value of any summary statistic from the stem plot. The median actually is 133 here, so all other statements are incorrect.

6. (a) The standard deviation measures the spread of the scores. If all the scores were the same, there is no spread, so the standard deviation will be 0.

7. (a) The effect of adding \$3,000 to each observation is to increase the first quartile by \$3,000 and to increase the third quartile by \$3,000. The difference, which is the interquartile range, is thus unchanged (the \$3,000s cancel).

8. (d) Imagine the scatterplot: We would draw drinking status, i.e., average number of drinks consumed per day for an individual, on the horizontal axis and the related score that represent the extent of the heart disease for this individual (that ranges from 0 to 100) on the vertical axis. Therefore, drinking status is the explanatory variable.

9. (a) Logan has the greater spread. Its high temperatures range from below 0 to around 100, whereas high temperatures in Berkeley only range from around 50 to around 100.

10. (a) Clearly 10 inches or less. The SD roughly represents an average departure from the average. It should be obvious that the average is around 65 inches. One half of the individuals are around 65 inches tall, so they have a very small departure from the average. The other half of the individuals measures either around 10 inches above or below the average — so their departure from the average is around 10 inches. Now we have 50 times 0 inches and 50 times 10 inches, so the average of these 100 departures from the average is much less than 10 inches.

11. (b) Only Histogram B is correct. We need 48 students in the class with 0 children (and not 42 or 43 as in Histogram A). Also, there are more students with 4 children than with 2 children, but Histogram A (incorrectly) indicates the same number for each class.

12. (b) There is no linear relationship between the two variables. But, there could be a perfect quadratic (i.e., curved) relationship that results in a correlation of 0.

13. (c) Calculate $-7 + (-5) + (-2) + 5 = -9$.

14. (d) Resort the data from smallest to largest:

    -7 -5 -2 3 5 7

    Then calculate $-5 + (-2) + 3 + 5 = 1$.

15. (b) We are summing up 6 times the same value (-2), so

$$-2 + (-2) + (-2) + (-2) + (-2) + (-2) = 6 \cdot (-2) = -12.$$